



Tenstorrent Blackhole p150

Also known as: Blackhole p150a · Blackhole p150b · p150a · p150b · TC-03003 · TC-03002

The Blackhole p150 is Tenstorrent's single-chip PCIe 5.0 x16 AI accelerator card: one Blackhole ASIC with 120 enabled Tensix cores, 180 MB of on-chip SRAM, 32 GB of GDDR6 at 512 GB/s and a 300 W board power, sold in an actively cooled p150a for workstations and a passively cooled p150b for server chassis. Its distinguishing feature is four 800G QSFP-DD ports on the card that link Blackhole cards directly to each other, so a multi-card mesh does not ride the host PCIe bus. At \$1,399 list it is the cheapest way to buy 32 GB of accelerator memory with a card-to-card fabric, but it runs Tenstorrent's own open-source stack (TT-Metalium, TT-NN), not CUDA, and its 512 GB/s memory bandwidth caps single-stream decode speed well below current NVIDIA cards.

664 TFLOPS Compute · BLOCKFP8, vendor claim; only precision Tenstorrent publishes	32 GB GDDR6 Memory · 512 GB/s, 16 GT/s	180 MB On-chip SRAM · 120 Tensix cores
4x QSFP-DD 800G Card-to-card links · passive, Blackhole-to-Blackhole only	300 W TBP Board power · 75 W idle	PCIe 5.0 x16 Host interface · no bifurcation

COMPUTE		01/05
ASIC	1x Tenstorrent Blackhole	
Tensix cores	120 · reduced from 140 by firmware v19.5.0 (early 2026), reported	
Big RISC-V cores	16 · general-purpose cores on the ASIC	
AI clock	Up to 1.35 GHz	
Peak compute	664 TFLOPS · BLOCKFP8, vendor claim; FP16/BF16 and FP4 figures not published	

MEMORY		02/05
On-chip SRAM	180 MB · about 1.5 MB per Tensix core	
Capacity	32 GB GDDR6	
Memory speed	16 GT/s	
Bandwidth	512 GB/s	
Sibling card	p100a: 28 GB at 448 GB/s · same ASIC, no QSFP-DD ports, \$999	
INTERCONNECT & FORM FACTOR		03/05
PCIe	Gen 5 x16 · no bifurcation; runs in narrower slots at possibly lower performance	
QSFP-DD ports	4x 800 Gbps · passive cables up to 2 m; connect only to other Blackhole cards	
Dimensions	42 x 270 x 111 mm · W x D x H, dual-slot	
p150a cooling	Active (fan) · part number TC-03003	
p150b cooling	Passive · part number TC-03002; needs chassis airflow	
POWER & ENVIRONMENT		04/05
Total board power	300 W	
Idle power	75 W	
External power	1x 12V-2x6 (12+4-pin) · ATX 3.1 certified PSU or better required	
Airflow	30 CFM minimum · at up to 35C inlet	
Operating temperature	10C to 35C	
HOST & SOFTWARE		05/05
Supported OS	Ubuntu 22.04 · per the card documentation	
Minimum host	x86_64 CPU, 64 GB RAM, 100 GB storage · 2 TB or more storage recommended	
Software stack	TT-Metalium / TT-NN · open source; not CUDA	
Model support	Per-model bring-up tables · Blackhole coverage still maturing behind Wormhole, reported	

PROCUREMENT

Typical lead time

Listed in stock on Tenstorrent's store, shipping in about 2 weeks plus 2 to 5 weeks transit, so plan on 4 to 7 weeks to delivery from a direct order.

Pricing

\$1,399 list for either the p150a or the p150b on Tenstorrent's store. The p100a (28 GB, no QSFP-DD) lists at \$999. A QSFP-DD 800G cable lists at \$200; buying two or more p150 cards gets 50% off one cable.

- Sold direct from Tenstorrent at a published list price, not allocation-gated or quote-only the way NVIDIA datacenter parts are.
- Pick the p150a for a workstation with no ducted airflow and the p150b only for a server chassis that delivers at least 30 CFM across the card.
- Each card needs a native 12V-2x6 cable from an ATX 3.1 supply; older PSUs with adapters are not supported, which rules out a lot of reused workstation chassis.
- Budget the QSFP-DD cables up front if you plan a multi-card mesh: the card-to-card fabric is the reason to pay \$400 more than the p100a.

FIELD NOTES

- 512 GB/s of GDDR6 is the decode ceiling. For single-user token generation a used RTX 3090 (936 GB/s) can still be faster, reported; the p150's case is memory per dollar and the on-card mesh, not raw single-stream speed.
- The 664 TFLOPS figure is BLOCKFP8, Tenstorrent's block floating-point format, and is the only precision the vendor publishes. Do not compare it directly with NVIDIA dense FP8 or FP4 numbers without a measured benchmark on your model.
- The four QSFP-DD ports are the real differentiator: cards talk to each other over 800G links without crossing host PCIe, so a four-card box scales tensor-parallel work better than four PCIe-only cards would. They only connect to other Blackhole cards; they are not a general Ethernet NIC.
- Active Tensix cores dropped from 140 to 120 with firmware v19.5.0 and the published peak fell from roughly 774 to 664 TFLOPS; Tenstorrent says typical workloads lose only about 1 to 2%, reported. Check which firmware any benchmark you read was run on.
- This is not a CUDA part. Every model needs a TT-Metalium/TT-NN bring-up, and published Blackhole model coverage trails Wormhole. Treat throughput claims as preliminary until you run your own model on the current release.

QUESTIONS WE GET ON THIS PART

What is the difference between the p150a and the p150b?

Cooling only. Both have the same Blackhole ASIC, 32 GB GDDR6 at 512 GB/s, four 800G QSFP-DD ports and 300 W board power. The p150a has an active fan for workstations; the p150b is passive and needs at least 30 CFM of chassis airflow in a server. Both list at \$1,399.

How much does the Blackhole p150 cost?

\$1,399 list from Tenstorrent's own store for either variant. The p100a, which has 28 GB and no QSFP-DD ports, is \$999. QSFP-DD 800G cables are \$200 each, with 50% off one cable when you buy two or more p150 cards.

Does the p150 run CUDA or PyTorch models directly?

No CUDA. It runs Tenstorrent's open-source stack (TT-Metalium and the TT-NN library), and each model needs support in that stack. Check Tenstorrent's model bring-up tables for your exact model and software release before buying.

How does the p150 compare with an NVIDIA consumer GPU for LLM inference?

It has more memory than most consumer cards at its price (32 GB) and a card-to-card fabric they lack, but 512 GB/s of memory bandwidth is lower than an RTX 3090 or 5090, so single-stream decode is slower. Its strength is scaling across several cards over the QSFP-DD mesh; independent benchmarks are still scarce.

REFERENCES

- [1] <https://tenstorrent.com/hardware/blackhole>
- [2] <https://docs.tenstorrent.com/aibs/blackhole/index.html>
- [3] <https://www.profesionalreview.com/2026/02/05/blackhole-p150-tenstorrent-aceleradoras-ia/>
- [4] <https://www.spheron.network/blog/tenstorrent-vs-nvidia-open-source-ai-hardware/>
- [5] https://dev.to/jovan_chan_9500711396d4e6/qualcomms-10b-tenstorrent-bid-what-risc-v-ai-cards-mean-for-home-labs-in-2026-ecd